

J•T•L•A

The Journal of Technology, Learning, and Assessment

Volume 6, Number 9 · June 2008

Does it Matter  
if I Take My Mathematics Test  
on Computer?  
A Second Empirical Study of  
Mode Effects in NAEP

Randy Elliot Bennett, James Braswell,  
Andreas Oranje, Brent Sandene,  
Bruce Kaplan, & Fred Yan

[www.jtla.org](http://www.jtla.org)

A publication of the Technology and Assessment Study Collaborative  
Caroline A. & Peter S. Lynch School of Education, Boston College

## Does it Matter if I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP

Randy Elliot Bennett, James Braswell, Andreas Oranje, Brent Sandene,  
Bruce Kaplan, & Fred Yan

Editor: Michael Russell  
russelmh@bc.edu  
Technology and Assessment Study Collaborative  
Lynch School of Education, Boston College  
Chestnut Hill, MA 02467

Copy Editor: Jennifer Higgins  
Design: Thomas Hoffmann  
Layout: Aimee Levy

JT LA is a free on-line journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

Copyright ©2008 by the Journal of Technology, Learning, and Assessment  
(ISSN 1540-2525).

Permission is hereby granted to copy any article provided that the Journal of Technology, Learning, and Assessment is credited and copies are not sold.

---

### Preferred citation:

Bennett, R.E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., Yan, F. (2008). Does it Matter if I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP. *Journal of Technology, Learning, and Assessment*, 6(9). Retrieved [date] from <http://www.jtla.org>.

# J·T·L·A

## **Abstract:**

This article describes selected results from the 2001 Math Online (MOL) study, one of three field investigations sponsored by the National Center for Education Statistics (NCES) to explore the use of new technology in NAEP. Of particular interest in the MOL study was the comparability of scores from paper- and computer-based tests. A nationally representative sample of eighth-grade students was administered a computer-based mathematics test and a test of computer facility, among other measures. In addition, a randomly parallel group of students was administered a paper-based test containing the same math items as the computer-based test. Results showed that the computer-based mathematics test was significantly harder statistically than the paper-based test. In addition, computer facility predicted online mathematics test performance after controlling for performance on a paper-based mathematics test, suggesting that degree of familiarity with computers may matter when taking a computer-based mathematics test in NAEP.

# Does it Matter if I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP

Randy Elliot Bennett, James Braswell, Andreas Oranje, Brent Sandene, Bruce Kaplan, & Fred Yan  
*Educational Testing Service*

## Introduction

In 2006–2007, 23 states were reported to offer computer-based assessments to measure achievement in U.S. schools (Bausell & Klemick, 2007). One state, Virginia, offered online tests in over a dozen subjects, administering approximately 1.5 million such assessments that school year (R. Triscari, personal communication, 6/19/07).

Projecting the emergence of electronic delivery, the National Center for Education Statistics (NCES) in 1999 commissioned three field studies to investigate the implications of computerized testing for the National Assessment of Educational Progress (NAEP). NAEP, called the “Nation’s Report Card,” periodically evaluates what fourth-, eighth-, and twelfth-grade students know and can do in a variety of school subjects. The field studies commissioned by NCES were conducted in mathematics, writing, and problem solving with technology. Results of the Writing Online study were reported in Horkay, Bennett, Allen, Kaplan, and Yan (2006), and the outcomes of the Technology-Rich Environments study can be found in Bennett, Persky, Weiss, and Jenkins (2007). The current paper reports selected results from the Mathematics Online (MOL) study, in particular those results concerning the comparability of scores across delivery modes.

Comparability is important because if delivery mode affects scores, NAEP’s ability to draw valid conclusions from test results may be reduced:

- If results are to be compared over time and the delivery mode has changed from paper to computer;
- If results are to be aggregated across individuals when some individuals have taken the test on paper and others have taken it on computer (especially if the assignment to modes was not voluntary); or

- If groups taking the test on computer are to be compared with one another and computer delivery affects one group more than another (such that between-group score differences become larger or smaller than they were for paper assessment).

At the K–12 level, Wang, Jiao, Young, Brooks, and Olson (2007) synthesized much of the available research through a meta-analysis of 38 effect sizes, finding no significant difference in mean performance between computer and paper mathematics tests. The available studies, however, were limited in important ways, including that most were unpublished and presumably not peer-reviewed, most used multiple-choice items only, the majority of the effects came from three investigations, and the representativeness of examinee samples was not considered.

Given the relatively small number of published studies, it is not surprising that results for population groups are rare. Among the published, peer-reviewed studies providing population-group results is that of Poggio Glasnapp, Yang, and Poggio (2005). These investigators reported findings from a sample of 644 grade 7 students who volunteered to be tested on a state mathematics assessment in both paper and computer modes. The main effect for test mode was not significant and no significant interactions of test mode with gender, socio-economic status, or academic placement (special education, general education, gifted education) were detected.

Also rare are published studies of the effect of computer familiarity on mathematics test performance. Russell (1999) assessed several small groups of students from two local schools, including one group randomly assigned to take six open-ended math items on computer and another group to take the same items on paper. All students were also administered a test of keyboarding skill. Russell found that, compared to a paper test, taking a constructed-response mathematics test on computer had a negative effect on scores but that this effect moderated as keyboarding skill increased.

In the current study, three main questions were addressed:

- Do students perform differently across modes at the total score and item level?
- Does mode differentially affect the overall performance of particular NAEP reporting groups (e.g., those categorized by gender or race/ethnicity)?
- Does computer familiarity appear to have an impact on online test performance?

## Method

### Participants

The target population consisted of eighth-grade students enrolled in public and private elementary and secondary schools in the United States.<sup>1</sup> A nationally representative, multi-stage, probability sample was selected. The procedure and sample are described briefly here with additional detail given in Appendix A.

In the first stage, the primary sampling units (PSUs) were counties or groups of counties. Middle and secondary schools were the sampling units in the second stage. In the third stage, schools were assigned to testing conditions. Because it would be costly to transport computers to a school to test only a few students, all schools were assigned to take part in both computer and paper-and-pencil conditions, with the exception of two very small schools that were assigned to administer paper-and-pencil only. Finally, in the fourth stage, students were randomly selected. In those schools selected to administer under both testing conditions, the selected students were assigned randomly to the online or paper-and-pencil forms. For all schools, students in the paper-and-pencil condition were assigned randomly to one of three parallel forms, only one of which was used in the analyses reported in this paper.

Students were tested in April and May of 2001. Those assigned to the online condition took the MOL test on school computers via the Internet or on disconnected NAEP laptops brought into schools. Sixty-two percent of students were assessed on laptop computers. All administrations, whether paper or online, were proctored by NAEP staff.

One hundred ten of the 129 sampled schools (87 percent) participated in the online condition and 108 of 131 sampled schools (83 percent) took part in the paper condition.<sup>2</sup> Of the 1,297 students sampled for the online condition, 1,072 students participated (84 percent). Of these 1,072 students, 56 were nonrespondents because of technology problems, reducing the tested sample to 1,016 participants. In the paper condition, 954 of 1,680 sampled students participated (83 percent). On average, 9 eighth-grade students per school were assessed on computer and approximately 26 were tested on paper.

Table 1 gives the weighted percentages of students by gender and race/ethnicity for each study condition. Also included for comparison purposes are the analogous percentages for the much larger, nationally representative sample participating in the 2001 main NAEP mathematics assessment (Braswell, Lutkus, Grigg, Santapau, Lim, & Johnson, 2001). As the table indicates, the study samples are generally comparable to one another and to the 2000 main NAEP nationally representative sample.

**Table 1: Percentages of Students by Gender and Race/Ethnicity for Study Conditions and for the 2000 Main NAEP Mathematics Assessment**

Group	MOL (N = 1,016)	P&P (N = 954)	2000 Main NAEP (N = 15,694)
Gender			
Male	49	51	51
Female	51	49	49
Race/Ethnicity			
White	67	66	67
Black	13	14	13
Hispanic	14	14	14
Asian American/Pacific Islander	5	4	4
American Indian	1	1	2

NOTE: MOL=Math Online. P&P=Paper and Pencil. Race/ethnicity data are based on student questionnaire responses. Gender data are based on school records.

## Instruments

All study participants took:

- *a paper-and-pencil block of mathematics questions*, administered first. The paper-and-pencil block contained 20 multiple-choice items from the NAEP 2000 mathematics assessment. The block was used for scaling purposes and also as a covariate in selected analyses.
- *a background questionnaire* to gather information about demographics and computer experience, presented last. The background questionnaire contained 30 questions with a 20-minute time limit.

After the initial math paper block, students taking the *computer-based* test (hereinafter referred to as MOL) received:

- *an online tutorial* in how to use the computer to complete the test. The online tutorial included instruction and practice in clicking on choices, clicking to shade or darken regions, moving back and forth between screens, correcting errors, and typing answers and explanations. The tutorial also had embedded tasks to provide a measure of the student's computer skill. The tutorial was split into two portions: a basic portion that preceded the test and a calculator portion that preceded the third test section. (The tutorials can be viewed at <http://nces.ed.gov/nationsreportcard/studies/tbatutorial.asp#mol>.)
- *online mathematics questions*, drawn from the existing NAEP item inventory and presented in three sections. Students were given paper to use for scratch work in answering these questions. There were 26 questions: 16 multiple choice, 8 short constructed response, and 2 extended constructed response. The time allowed for each section was 15 minutes and the number of questions per section was 10, 9, and 7, respectively. The third section permitted use of an on screen, scientific calculator modeled in layout and functionality after the handheld calculator used in main NAEP mathematics assessments. The on screen calculator was available to students throughout the test section, though it was intended to be helpful for solving only some of the items in that section.

After the initial paper block, students taking the *paper* test took a form, referred to here as "P&P" that contained the same three sections of 26 mathematics questions as the online test, with the same time limits. The third section of this test permitted the use of a handheld, scientific calculator provided by NAEP administrators.

Because the paper and computer tests were comprised of the “same” items, the two forms were putatively identical in their mathematical content. However, because those items were originally written for paper presentation, some items needed to be changed for computer delivery. The overwhelming majority of items were changed minimally in their presentation (e.g., on paper, figures typically were placed above the question text, whereas on computer they were placed to the left of the text; wording for some items was changed from “make a mark ...” to “click on...”). Four items needed to be adapted more noticeably. Several of these instances are discussed in the Results section of this paper.

Table 2 provides an overview of the instruments and student samples. From the table, it should be clear that different, randomly parallel student groups took the same test, one group on computer and the other group on paper. In addition, both groups took a common initial block of items on paper. Performance on the initial paper block provides a convenient mechanism for checking the equivalence of the samples. The raw-score means were 12.4 and 12.3, for the MOL and P&P samples, respectively.

**Table 2: Instruments Administered to Each Student Sample**

Sequence	MOL (N = 1,016)	P&P (N = 954)
1	Initial paper block (20 items)	Initial paper block (20 items)
2	Online tutorial	†
3	Online test (26 items) with embedded calculator tutorial	Paper test (P&P) (26 items)
4	Background questions (30 items)	Background questions (30 items)

† Not applicable

NOTE: MOL=Math Online. P&P=Paper and Pencil .

## Procedure

### Constructed-response Scoring

The test administered to each sample contained 10 constructed-response questions. A team of trained raters scored responses to these items. Raters used the rubrics and sample answers that had been developed for the items when those items were used in NAEP paper assessments. Where needed, supplemental training responses were printed from the online versions of the questions.

A single team scored both the online and the paper responses to each item. Responses written in test booklets were scored on paper; those completed on computer were presented to raters for scoring on computer.

A random sample of approximately 25 percent of the responses was double-scored to compute inter-rater reliability. The median exact agreement was 95 percent for P&P (range = 80 percent to 99 percent) and 94 percent for MOL (range = 84 percent to 98 percent). (Appendix B presents results for each item.)

### Scaling and Proficiency Estimation

To scale items and estimate examinee proficiencies, the study used essentially the same process employed for NAEP assessments. (See Allen, Donoghue, & Schoeps, 2001, for complete details on these NAEP technical procedures.) Departures from the procedures typically used for NAEP assessments are noted, as appropriate.

Calibration was conducted with the 3-parameter logistic model for multiple-choice items and the generalized partial credit model (Muraki, 1992) for constructed-response items, as implemented in the NAEP version of Parscale. Using these IRT models, the item parameters for the initial paper block, MOL, and the paper forms were estimated together (45 questions in all). (One item was omitted from the analysis because it introduced difficulties in obtaining a satisfactory scaling solution.) This univariate calibration step was repeated with several model variations for use in different analyses. For example, to facilitate the study of total-score mode effects, the calibration was conducted with item parameters constrained to be equal across MOL and the P&P form. For item-level comparisons, however, the calibration was conducted with parameters permitted to vary across the two testing modes. For such calibrations, the initial-paper-block items were constrained to be equal across examinee groups, thereby defining a common scale on which MOL and the paper form could be compared. This constraint assumed that the initial-paper-block, MOL, and the paper form each measured the same unidimensional skill.

As is routine in NAEP, examine proficiencies were generated from student demographic information (to remove bias from the estimation of group performance), the item-response-model assumptions, the item parameters estimated in the calibration step, and item responses to the MOL or paper test. Scores were placed on an arbitrary scale with a mean of 200, a standard deviation of 30, and a range from 0 to 400.

## Results

### Performance Differences Across Delivery Modes

The analysis of performance differences across delivery modes centers on two questions:

- Are there differences in mean scores?
- Are there differences in item functioning?

#### Differences in Mean Scores

To address the first question, the mean scale scores for MOL and P&P were compared. For this analysis, mean scores were generated from a scaling in which the item parameters for each mode were constrained to be equal, thereby forcing mode differences into the total scores. For MOL, the mean scale score was 198, whereas for P&P it was 202. This difference was statistically significant ( $t_{(52)} = -2.26, p < .05$ ).<sup>3</sup> In terms of practical importance, the difference of .14 standard deviations is less than the .2 minimum for “small” effects suggested by Cohen (1988).

In addition to differences in central tendency, differences in the spread of the two distributions are of interest. The standard deviation of the MOL score distribution was 32, whereas for P&P it was 27, suggesting greater variability in the computer-based scores.

#### Differences in Item Functioning

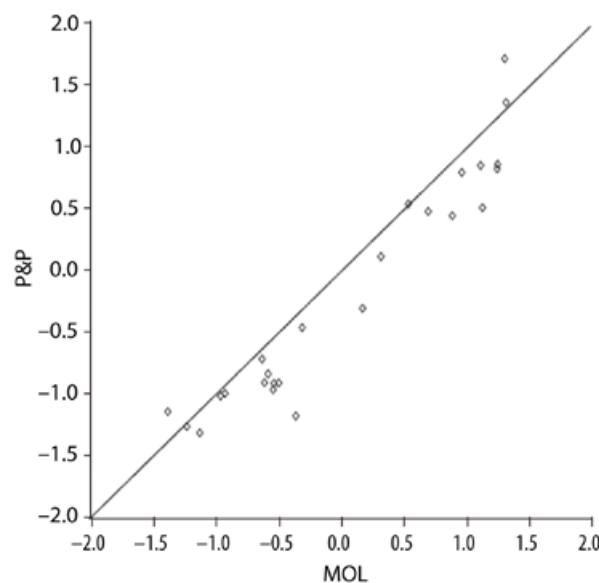
In the IRT framework used in this study, item functioning is characterized by three parameters (difficulty  $b$ , discrimination  $a$ , and guessing  $c$ ) for multiple-choice items. For polytomously scored, constructed-response items, it is common practice not to model guessing, since the probability of obtaining a correct response by chance is considered to be negligible. Also, for such items, the probabilities associated with more than two response categories have to be modeled. Therefore, in addition to a discrimination parameter  $a$ , several item step parameters are used that indicate the regions of the ability scale with which those item response categories are most likely to be associated. An overall location parameter  $b$  is also estimated.

To address the study question related to item functioning, mode differences in difficulty and discrimination were focused upon because these parameters are the main drivers of potential overall differences between presentation modes. For each of the items, IRT  $b$  (difficulty/location) and  $a$  (discrimination) parameters were estimated as part of scaling, using the examinee response data from the two presentation modes and allowing item parameter estimates to vary across those modes. This analysis helps to identify whether the mode effects that were observed at the total score level are linked to relatively uniform differences in item functioning or, alternatively, are the result of a small number of outliers.

The IRT  $b$  parameter is considered first. This parameter positions the item on the ability scale at the point where the probability of a correct response is .5. The parameter is commonly estimated to range from -2.0 to 2.0. Items with higher  $b$  values are more difficult.

Figure 1 presents the scatter plot of the IRT  $b$  values for the 25 paper-administered items against the  $b$  values for the same 25 MOL (Math Online) items. Two results stand out. First, the relationship of the estimated parameters to one another is almost identical across modes: the product-moment correlation is .96. Second, the preponderance of items falls on the MOL side of the identity line, suggesting that items presented on computer were generally more difficult than the same items on paper.<sup>4</sup>

**Figure 1:** Comparison of IRT  $b$  parameter estimates for items presented on MOL and on paper, each to a different sample of eighth-grade students



NOTE: MOL=Math Online. P&P=Paper and Pencil.

Table 3 (next page) shows the IRT  $b$  parameter estimates for two subclasses of item: 21 items needing minimal change to render on computer and four items needing considerable change for computer rendering. Items are ordered in the table by the size of the cross-mode difficulty difference, beginning with items observed in the sample to be easier on computer. Also shown in the left-most columns are the item administration order and the format. Three item formats were used: multiple-choice (MC), short constructed-response (SCR), and extended constructed-response (ECR). SCR questions were scored on either a 2- or 3-point scale, while ECRs were scored on a 5-point scale. The response requirements of SCR and ECR questions included such things as clicking on line segments to create a geometric figure, making a numeric entry, clicking on a number line to specify a value, and entering text to justify an associated response.

**Table 3: IRT  $b$  parameter estimates for items presented on computer and on paper, by extent of change required for computer**

Item #	Format	Estimated $b$		Computer–Paper Difference
		Computer	Paper	
<b><i>Items Needing Minimal Change to Render on Computer</i></b>				
11	MC	−1.38	−1.14	−.25
8	MC	1.32	1.37	−.05
1	MC	.54	.55	−.01
18	MC	−1.23	−1.25	.02
12	MC	−.96	−1.01	.05
3	SCR	−.93	−.99	.06
2	SCR	−.63	−.72	.09
5	MC	−.31	−.46	.15
23	MC	.97	.80	.17
20	MC	−1.13	−1.31	.18
14	MC	.32	.13	.20
22	MC	.70	.49	.21
4	MC	−.58	−.84	.26
25	MC	1.12	.85	.27
6	MC	1.26	.87	.39
19	SCR	−.50	−.90	.41
24	MC	1.25	.83	.41
7	SCR	.89	.46	.43
21	MC	.18	−.30	.48
26	ECR	1.14	.52	.62
13	SCR	−.36	−1.16	.81
<b><i>Items Needing Considerable Change to Render on Computer</i></b>				
10	ECR	1.31	1.73	−.42
17	SCR	−.60	−.90	.30
15	SCR	−.53	−.91	.38
16	SCR	−.54	−.96	.41

NOTE: MC=multiple choice. SCR=short constructed-response. ECR=extended constructed-response. For polytomous items, the estimated  $b$  is the item difficulty/location following the parameterization of Muraki (1992).

Taken across all 25 items shown in the table, the mean of the differences was equal to .22 logits (range = -.25 to .81). Because positive and negative differences can cancel each other out, the mean of the absolute values of the differences was also calculated. This value equaled .28 logits.<sup>5</sup>

As the bottom section of Table 3 indicates, three of the four items requiring considerable change for computer rendering appeared to be more difficult than their paper counterparts, whereas the fourth item appeared to be easier. Taken across all four items, the mean differences in logits for the considerably changed vs. minimally changed items were .17 vs. .23, respectively, and the mean absolute differences were .38 vs. .26.

The three considerably changed items that appeared harder on computer were implemented quite differently compared with their paper renderings. For each of these items (#15–17), the general task was to determine the value of a point on a number line. On the paper test, the examinee only needed to write a value on the number line in the space provided. On the computer test, the student first had to choose the appropriate answer-template (a whole number, decimal, fraction, or mixed number), and then type the answer into that template.

As the table suggests, change in presentation was associated with response format: the questions needing considerable change were all constructed response. Thus, it is not surprising that re-classifying the data by item format also suggests an impact on difficulty. On average, the discrepancies were about twice as large for constructed-response questions as for multiple-choice items: the mean difference for constructed-response was .31 vs. .16 logits for multiple-choice, and the mean absolute differences were .39 and .20, respectively.

Finally, items were classified by whether or not a calculator was present. (Recall that a handheld, scientific calculator was made available for section three of P&P, and an online scientific calculator modeled on the handheld one was available for that same section in MOL.) Since the calculator was only present for items in the final section of the test (i.e., items 20–26), it should be noted that this comparison confounds position with difficulty. The mean difference between paper and computer presentation for the seven calculator-present items was .33 logits and the mean absolute difference was also .33. For the 18 items where the calculator was not available, the comparable figures were .18 and .26, suggesting the possibility that the presence of a calculator might increase mode differences somewhat.

The IRT  $a$  parameter describes the discrimination of an item, and is commonly considered to be the analog of the classical item-total correlation. Items with lower  $a$  values do not differentiate between examinees at particular points on the ability scale as well as items with higher values.

Table 4 (next page) gives the discrimination estimates for each item in computer- and paper-based administrations, and the difference between the estimates. As the table indicates, 16 of the 25 items appeared to be more discriminating on paper than on computer. However, across all 25 items, the mean of the discrimination differences was  $-.04$  and the mean of the absolute differences was  $.13$ , suggesting minimal effects. Also, the parameter estimates were highly related across modes ( $r = .86$ ), though not as highly as the difficulty estimates.

**Table 4: IRT  $\alpha$  parameter estimates for items presented on computer and on paper, by extent of change required for computer**

Item #	Format	Estimated $\alpha$		Computer–Paper Difference
		Computer	Paper	
<b><i>Items Needing Minimal Change to Render on Computer</i></b>				
25	MC	.86	1.35	-.50
22	MC	.83	1.13	-.30
18	MC	.86	1.17	-.30
4	MC	.74	.98	-.25
7	SCR	.70	.88	-.18
1	MC	1.01	1.16	-.16
6	MC	1.22	1.31	-.10
12	MC	.76	.84	-.09
11	MC	.92	.99	-.06
5	MC	.58	.63	-.05
2	SCR	.62	.66	-.04
13	SCR	.39	.43	-.04
20	MC	.79	.81	-.02
19	SCR	.47	.49	-.02
14	MC	1.37	1.39	-.01
3	SCR	.42	.42	#
26	ECR	.78	.77	.01
21	MC	.88	.80	.09
8	MC	1.05	.91	.14
24	MC	1.19	1.03	.17
23	MC	1.13	.93	.20
<b><i>Items Needing Considerable Change to Render on Computer</i></b>				
17	SCR	1.49	1.60	-.11
16	SCR	1.44	1.32	.12
15	SCR	1.45	1.27	.17
10	ECR	.61	.36	.25

NOTE: MC=multiple choice. SCR=short constructed-response. ECR=extended constructed-response.

# The estimate rounds to zero.

Items needing considerable change for computer presentation did not differ much from items needing minimal change in their power to discriminate as measured by IRT  $\alpha$  parameter estimates. The mean difference for the changed items was .11 and for the unchanged items -.07. The mean absolute differences were .16 versus .13.

### Population Group Performance

To investigate whether NAEP reporting groups were differentially affected by computer presentation, mean performance on the computer-presented test was compared with mean performance on the paper form (P&P). Comparisons were made for groups categorized by gender, race/ethnicity, parents' education level, school location, region of the country, and school type. Within each such group (e.g., males), the difference between the MOL and P&P mean scores was evaluated using an independent-samples  $t$ -test, correcting for chance via the false discovery rate (FDR) procedure (Benjamini & Hochberg, 1995) for the number of tests run for the category (e.g., two  $t$ -tests for gender, one for males and one for females).

Table 5 (next page) gives the means and standard deviations for each group. Because the sample sizes for some groups were quite small, differences may not always be statistically significant even if they are seemingly large. It is not possible to distinguish for these instances whether the apparent difference is a true reflection of the population performance or, alternatively, an artifact of sample selection. For the groups examined, only one statistically significant difference was detected: Students reporting that at least one of their parents graduated from college performed better on P&P than a comparable group taking the same test on computer ( $t,64 = -2.73, p < .05$ ). For this group, the difference in mean scores was 6 points, or an effect of .21 standard deviation units, which would be characterized as "small" in Cohen's (1988) classification. Also, as the table shows, for this group as for the total group, MOL scores appeared to be more variable than the P&P ones.

**Table 5: Means and standard deviations by NAEP reporting group for MOL and the paper-and-pencil test form**

Category and Group	Sample Size		Mean (SD)	
	MOL	P&P	MOL	P&P
Gender				
Male	495	491	199 (32)	203 (27)
Female	521	463	197 (33)	201 (27)
Race/Ethnicity				
White	613	566	206 (30)	209 (24)
Black	126	135	176 (31)	185 (25)
Hispanic	224	202	178 (29)	185 (27)
Parents' Education Level				
Did not finish HS	73	63	178 (30)	182 (24)
Graduated HS	208	181	189 (31)	194 (24)
Some education after HS	179	171	199 (30)	200 (27)
Graduated college	471	422	205 (32)	211 (25)*
Unknown	76	108	186 (32)	191 (28)
School Location				
Central city	330	319	188 (34)	192 (30)
Urban/fringe	455	406	204 (31)	207 (25)
Rural	231	229	198 (31)	205 (24)
Region				
Northeast	185	140	202 (35)	206 (25)
Southeast	298	272	187 (32)	194 (25)
Central	191	204	208 (28)	210 (25)
West	342	338	198 (32)	201 (29)
School Type				
Public	945	885	197 (33)	201 (27)
Nonpublic	71	69	205 (27)	214 (21)
Total Group	1,016	954	198 (32)	202 (27)*

\*  $p < .05$  for the test of the differences in means between MOL and P&P.

NOTE: MOL=Math Online. P&P=Paper and Pencil.

## Performance as a Function of Computer Experience

In considering the potential impact of computer familiarity on computer-based test performance, it may be sensible first to ask how familiar eighth-grade students are with computers. The current study administered background questions selected from ones used in the (paper) NAEP 2001 history and geography assessments. Responses to these questions (collapsed across the MOL and paper conditions) suggested that most students had some familiarity with computers. Eighty-three percent of students indicated that there was a computer at home and 74 percent said they used it to access the Internet. In addition, a majority said that at least once a week, they used a computer outside of school (83 percent). A majority said they used one at school at least once a week (56 percent). Most students reported employing a computer to find information on the Internet for school (93 percent) or personal use (88 percent), to play games (88 percent), to write (86 percent), to communicate via e-mail (81 percent), to look up information on a CD (80 percent), to chat (76 percent), to make drawings (72 percent), or to make tables, charts, or graphs (59 percent). Finally 44 percent said they employed a computer at school for mathematics at least once a week.

To determine whether familiarity with computers might affect online test performance, the relationship between computer familiarity and performance on the MOL test was examined. This analysis was conducted only for the total group of students taking the online test.

For purposes of this study, computer familiarity was conceived as having three components: computer experience, input accuracy, and input speed. A minimal level on each component should, in theory, be present before a student can effectively take an online test, especially one that includes constructed-response questions. For example, some amount of previous computer experience should allow quicker adaptation to the test's navigational and input procedures, which in the MOL test were designed to follow common software conventions. Likewise, input accuracy should be necessary for the student's intended answer to be recorded correctly. Finally, reasonable input speed is required because the MOL test gives students a limited time for completion; time lost to input that is accurate but slow might introduce irrelevant variance into test performance. In fact, such an effect for speed in online mathematics test performance has been found in at least one previous comparability study (Russell, 1999).

To measure the first component of familiarity, computer experience, a scale was created based on students' responses to background questions adapted from the NAEP 2001 history and geography assessments. The rationale for using background questions as a measure of experience was two-fold. First, NAEP employs such questions operationally to document

the extent and type of computer use among students. Second, very similar background questions have been used in other comparability studies as surrogates for computer proficiency (e.g., Taylor, Jamieson, Eignor, & Kirsch, 1998).

The score for the computer experience measure was the simple sum of the responses to each question, ranging from 0–40. (See Appendix C for the questions that were included in the computer experience measure.) While other question-aggregation rules are possible, this scheme was judged reasonable given research suggesting that different aggregation rules often produce similar results (Stanley & Wang, 1970).

The second and third components of computer familiarity, input accuracy and input speed, were measured using tasks embedded in the MOL tutorials (available at <http://nces.ed.gov/nationsreportcard/studies/tbatutorial.asp#mol>). Coming from the tutorial, the tasks were essentially the same mechanical ones that students needed to perform in taking the MOL test.

Table 6 (next page) shows the tasks included in the accuracy and speed measures. The accuracy scale range was 0–17 and the speed scale range was 0–22.

**Table 6: Components of the Input-Skill measure**

<b>Variable</b>	<b>Maximum Points</b>
<b>Accuracy</b>	
Typing and editing	
Accuracy typing a brief given passage	2
Accuracy inserting a word	2
Accuracy changing a word	2
Navigating the test	
Accuracy pointing and clicking with the mouse	2
Accuracy scrolling	2
Accuracy clicking on the "Next" icon	2
Accuracy clicking on the "Previous" icon	2
Entering responses	
Accuracy filling in a mixed number	2
Using the calculator	
Accuracy in performing a given operation	1
<b>Speed</b>	
Typing and editing	
Time to type a brief passage	2
Time to insert word	2
Time to change word	3
Navigating the test	
Time to point and click	3
Time to scroll	2
Time to click on "Next"	3
Time to click on "Previous"	2
Entering responses	
Time to fill in mixed number	2
Using the calculator	
Total time to complete the calculator tutorial	3

Coefficient alpha reliabilities for computer experience, input accuracy, and input speed were .78, .48, and .72, respectively. Correlations among the measures, and with the MOL test, are shown in Table 7. As can be seen, the correlations among the three computer-familiarity measures are generally quite a bit lower than the limit imposed by their reliability values.

**Table 7: Observed correlations among computer familiarity measures and with mathematics performance**

	Initial paper mathematics block	MOL test	Computer experience	Input accuracy
MOL test	.72			
Computer experience	.13	.21		
Input accuracy	.35	.39	.12	
Input speed	.44	.54	.31	.26

NOTE: All values are unweighted.

To explore the relationship between computer familiarity and performance in the computer-based test, an ordinary least-squares multiple regression was executed. The goal of this analysis was to determine if computer familiarity predicted performance on the computer-based test after controlling for mathematics skill as measured on paper. The independent variables were self-reported computer experience, input accuracy, input speed, and number-right raw score on the initial paper mathematics block, which served as a covariate. The dependent variable was the sum of the dichotomously scored and polytomously scored MOL test items. All three computer-experience variables were used because they are logically and empirically related to taking a mathematics test on computer, and not highly correlated with one another.

In conducting this regression analysis, simpler models were successively compared to more complex models so that, for example, the effect of the computer familiarity variables (entered as a block) on MOL score could be separated from the effect of the covariate on MOL score. Standard checks for residual outliers, multicollinearity, and influential observations were performed. The resulting full-factorial model accounted for 57 percent of the variance in predicting MOL score ( $F,8,539 = 165.92, p < .001$ ).

Table 8 (next page) gives the results for the main-effects model only because the interactions were not significant ( $F,4,539 = 0.73, p > .05$ ). After controlling for mathematics proficiency on the paper-based block, input accuracy and input speed significantly added statistically to the prediction of MOL score; self-reported computer experience did not add significantly statistically. In terms of the size of the effect, the initial paper block accounted for 49 percent of the variance in MOL scores. Adding

the computer familiarity variables to the model increased the variance accounted for in MOL scores to 57 percent.

**Table 8:** **Regression results for the effect of input skill and computer experience on MOL test raw score, controlling for paper mathematics proficiency**

Variable	Estimated regression coefficient	Standard error
Intercept	-15.78	2.327
Initial paper math block (covariate)	.87*	0.136
Input accuracy	.67*	0.131
Input speed	.37*	0.067
Computer experience	.05	0.025

\*  $p < .05$ , two-tailed t-test ( $df$ -range 3 to 12,  $t$ -range 1.86 to 6.36).

NOTE: The number of students included in the analysis was 1,011. A jackknife, replicate-weight procedure was used to compute the standard errors (see Allen, Donoghue, & Schoeps, 2001).

## Discussion

The Math Online study collected data from samples intended to represent the population of eighth-grade students in the United States. Students in more than 100 schools participated. The study addressed three main questions related to the comparability of computer and paper mathematics test scores:

- Do students perform differently across modes at the total score and item level?
- Does mode differentially affect the overall performance of particular NAEP reporting groups?
- Does computer familiarity appear to have an impact on online test performance?

With respect to performance, the mean scale score for eighth-graders taking the computer test was significantly lower statistically than that for a randomly parallel group taking the paper version of the same 25-item measure. In effect-size terms, however, the difference of .14 standard deviation units was not large enough to even be considered “small” in Cohen’s (1988) classification. At the item level, the difficulties for the computer test were generally greater (by an average of .22 logits on the IRT scale, or about .05 points on the proportion-correct scale), an expected finding given the direction of the differences in mean scores. Differences in item discrimination appeared to be negligible on average.

The second study question concerned the effect on the overall performance of NAEP reporting groups. The performance of selected groups was evaluated to see whether their mean scores differed on paper vs. computer versions of the same test. Separate comparisons were made by gender, race/ethnicity, parents' education level, region of the country, school location, and school type. Results showed that, for the NAEP reporting groups examined, mean performance generally was not differentially affected by electronic vs. paper delivery.

The last question dealt with the impact of computer familiarity on test performance. Students' responses to background questions suggested that the overwhelming majority of pupils used computers at home and at school. To determine if lack of computer familiarity affected online test performance, hands-on measures of input accuracy and input speed, and a measure of self-reported computer experience, were used to predict online test performance. After controlling for performance on a paper mathematics test, input speed and input accuracy predicted MOL score at statistically significant levels. The direction of the effect was such that the greater the student's computer familiarity, the higher was the predicted MOL score, suggesting that some students may have scored better on MOL than their equally mathematically proficient peers simply because the former students were more facile with computers. The increment in variance accounted for in MOL score after controlling for paper mathematics score was 8 percentage points. In evaluating this increment, note that admissions test scores like those from the SAT or ACT Assessment add only about 5 or 6 percentage points over high school grades in predicting first-year college grade-point-average (Burton & Ramist, 2001; Noble & Sawyer, 2002).

This result is similar in kind to that found by Russell (1999) and by Horkay et al. (2006). For mathematics, Russell's study included only constructed-response items given to approximately 200 eighth-grade students in Massachusetts. He found that, compared to taking a test on paper, taking the test on computer had a negative effect, which lessened as keyboarding skill increased. Horkay et al.'s study concerned the assessment of writing proficiency in NAEP. They presented essay tasks to national samples of eighth graders, finding that computer familiarity predicted online essay score after accounting for paper writing performance.

What causes the effects found in these studies? One possible contributing factor is the presence of constructed-response items, which can sometimes demand considerable computer skill. In the Russell study, all items required the student to key-enter at least a sentence of text. When asked what problems they had taking the mathematics test online, 30 percent of the students indicated they had difficulty typing. In the Horkay

et al. study, the keyboarding demands were considerably greater as students were asked to generate essays on computer.

In the present investigation, constructed-response items appeared to shift in difficulty more than multiple-choice items when presented on computer as compared with paper.<sup>6</sup> Constructed-response items also needed to be adapted more than multiple-choice items in order to be rendered on screen. These results suggest that, in moving paper mathematics items to computer, it may sometimes be harder to hold difficulty constant for constructed-response than for multiple-choice questions. The causes of that difficulty shift are not clear. The translation of constructed-response items to electronic delivery may introduce the need for computer skill in responding, may make it impossible for students to answer in alternative ways (e.g., diagrammatically), or may otherwise change the nature of what is being measured.

Two other factors that may have affected performance are technology problems encountered during test administration and the use of NAEP laptop computers. With respect to technology problems, approximately 11 percent of students were prevented from working through the tutorials and the test questions without interruption. Such interruptions might have occurred due to the loss of an Internet connection or to a hardware or software issue. In such cases, test administrators attempted to restart students where they had stopped or, if this was unsuccessful, returned them to the beginning of the test on the same or on an alternative machine. It is possible that after such interruptions, these students were less motivated and performed more poorly as a result. To evaluate the relationship between session interruption and performance, MOL score was regressed on test session status (interrupted vs. not interrupted), controlling for performance on the initial paper math block. This regression produced a statistically significant effect for session status ( $F, 1,35=12.43, p <.01$ ). However, the impact on scores appears to have been minimal. The effect's magnitude can be estimated by using the regression to predict what the MOL scores of students with interrupted sessions would have been had those interruptions not occurred. When the MOL mean for the total group is recalculated using predicted scores for students with interrupted sessions, and the actual scores of those with not-interrupted sessions, the sample mean increases marginally from 198 to 199.

The second factor, the use of NAEP laptops, is relevant because the majority of students took their tests on these computers, which would often have been less familiar than their school machines. To determine whether taking the test on a laptop was associated with student performance, MOL score was regressed onto computer type (school computer vs. NAEP laptop), with score on the initial paper math block serving

as a covariate. Computer type was a statistically significant predictor ( $F_{1,35}=82.54, p < .00$ ). An estimate of the effect of computer type can be gained by using the regression to predict what the MOL scores of students who took the test on laptop would have been had they taken it on desktop. This estimate needs to be regarded cautiously, however, because there may be other factors correlated with taking the test on laptop that would affect performance *regardless* of computer type (e.g., level of computer familiarity). When the MOL mean for the total group was recalculated using predicted scores for students taking the test on laptop, and the actual scores of those administered the test on desktop, the sample mean increased from 198 to 200. This increase in mean score likely overlaps with that of the increase predicted for students with interrupted sessions, as close to half of the students experiencing interrupted sessions took their tests on laptop computers. In any event, it seems that somewhat greater comparability between the computer and paper tests might have resulted from administering a larger proportion of the tests on school computers.

Several limitations should be considered in interpreting the results of this study. First, the data were collected in 2001, which from a technology perspective, was a long time ago. Since that time, students have become more comfortable with computers and especially with laptops. Additionally, laptop computers themselves have become easier to use. Screens have become larger and sharper, and keyboards easier to manipulate. Third, the technology has become more dependable, with technical problems being less likely to interrupt testing sessions. Finally, test developers have become more adept at designing computer assessments so that navigation, item presentation, and response entry are quicker to learn and simpler to accomplish. Taken together, these changes may have made computer and paper delivery more comparable over time.

A second limitation relates to the results for NAEP reporting groups. The sample sizes for these groups were often small, resulting in limited power to detect mode differences. Further, the impact of computer familiarity on performance was assessed only for the overall study sample, and not for reporting groups. It is possible that computer familiarity was associated with online math performance differently from one group to the next (e.g., for White vs. Black vs. Hispanic students). Future research should evaluate this possibility.

Third, the methods used to evaluate the relationship between computer familiarity and online math performance were correlational. This fact means that other unmeasured variables associated with computer familiarity could be responsible for the detected effect. An elaboration on the design used in the current study would have been to administer a computer familiarity measure to the group taking the paper test. Then,

the relationship of computer familiarity to paper math performance, after controlling for score on the initial paper block, could have been evaluated. The absence of such a relationship in the paper delivery condition might have provided stronger evidence for linking differences in computer familiarity to online math performance.

Finally, some studies in the field of writing assessment have detected a presentation effect in scoring, where handwritten answers received higher grades than typed versions of the same responses (Powers, Fowles, Farnum, & Ramsey 1994; Powers & Farnum 1997; Russell & Tao, 2004a, 2004b). It is unclear if the same effect would occur for NAEP mathematics items, particularly since the responses in the present study involved much less text than in essay examinations. In the current study, five of the ten constructed-response items required only simple numeric entry or clicking on hot spots, while the remaining five questions entailed text of no more than a few sentences. Further research might examine whether the MOL mode effect is partly due to reader bias by transcribing a sample of responses from each mode to the other, and having different readers grade subsets of the transcribed and original versions blindly.

## References

Allen, N.L., Donoghue, J.R., and Schoeps, T.L. (2001). *The NAEP 1998 technical report* (NCES 2001-509). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Bausell, C.V., & Klemick, E. (2007, March 29). Tracking US trends. *Education Week*, 26(30), 42-44.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289-300.

Bennett, R.E., Persky, H., Weiss, A.R., & Jenkins, F. (2007). Problem solving in technology-rich environments: A report from the NAEP technology-based assessment project (NCES 2007-466). Washington, DC: National Center for Education Statistics, US Department of Education. Retrieved June 12, 2008 from <http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007466>

Braswell, J.S., Lutkus, A.D., Grigg, W.S., Santapau, S.L., Tay-Lim, B.S.-H., & Johnson, M.S. *The nation's report card: Mathematics 2000* (NCES 2001-517). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Burton, N., & Ramist, L. *Predicting success in college: SAT studies of classes graduating since 1980* (Research Report No. 2001-2). New York: College Board. Retrieved January 31, 2006 from [http://www.collegeboard.com/research/pdf/rdreport200\\_3919.pdf](http://www.collegeboard.com/research/pdf/rdreport200_3919.pdf)

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2<sup>nd</sup> ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Horkay, N., Bennett, R.E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment*, 5(2). Retrieved July 29, 2007 from <http://escholarship.bc.edu/jtla/vol5/2/>.

Lapp, M.S., Grigg, W.S., and Tay-Lim, B.S. (2002). *The Nation's report card: U.S. history 2001*. Washington, DC: National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159-176.

NCS Pearson. (undated). *National Assessment of Educational Progress: 2000 report of processing and professional scoring activities, main and state NAEP*. Minneapolis, MN: Author.

Noble, J., & Sawyer, R. (2002). *Predicting different levels of academic success in college using high school GPA and ACT composite score* (ACT Research Report Series 2002-4). Iowa City: ACT. Retrieved January 31, 2006 from [http://www.act.org/research/reports/pdf/ACT\\_RR2002-4.pdf](http://www.act.org/research/reports/pdf/ACT_RR2002-4.pdf)

Poggio, J., Glasnapp, D.R., Yang, X., & Poggio, A.J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment*, 3(6). Retrieved July 5, 2007 from: <http://www.jtla.org>

Powers, D., and Farnum, M. (1997). *Effects of mode of presentation on essay scores* (RM 97-8). Princeton, NJ: Educational Testing Service.

Powers, D., Fowles, M., Farnum, M., and Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement*, 31(3): 220-233.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives*, 7(20). Retrieved August 5, 2003, from <http://epaa.asu.edu/epaa/v7n20/>.

Russell, M., & Tao, W. (2004-a). Effects of handwriting and computer-print on composition scores: A follow-up to Powers, Fowles, Farnum, & Ramsey. *Practical Assessment, Research and Evaluation*, 9(1). Retrieved July 10, 2005 from <http://PAREonline.net/getvn.asp?v=9&n=1>

Russell, M., & Tao, W. (2004-b). The influence of computer-print on rater scores. *Practical Assessment, Research and Evaluation*, 9(10). Retrieved July 10, 2005 from <http://PAREonline.net/getvn.asp?v=9&n=10>

Sandene, B., Bennett, R.E., Braswell, J., & Oranje, A. (2005). Online assessment in mathematics. In B. Sandene, N. Horkay, R.E. Bennett, N. Allen, J. Braswell, B. Kaplan, & A. Oranje (Eds.), *Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project* (NCES 2005-457). Washington, DC: National Center for Education Statistics, US Department of Education. Retrieved July 29, 2007 from <http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005457>.

Stanley, J.C., and Wang, M.D. (1970). Weighting test items and test-item options, an overview of the analytical and empirical literature. *Educational and Psychological Measurement*, 20: 21–35.

Taylor, C., Jamieson, J., Eignor, D., and Kirsch, I. (1998). *The Relationship between computer familiarity and performance on computer-based TOEFL® test tasks* (Report No. 61). Princeton, NJ: Educational Testing Service.

Wang, S., Jiao, H., Young, M.J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K–12 mathematics tests. *Educational and Psychological Measurement*, 67. Retrieved July 16, 2007 from <http://epm.sagepub.com/cgi/content/abstract/67/2/219>

Weiss, A.R., Lutkus, A.D., Hildebrant, B.S., and Johnson, M.S. (2002). *The nation's report card: Geography 2001* (NCES 2002–484). Washington, DC: National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education.

## Author Notes

James Braswell is at the American Institutes for Research (AIR). He was employed at ETS at the time this study was conducted.

This study was funded by the National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education under contract number ED-02-CO-0023.

We are indebted to many individuals for their help in completing this study. The full list can be found in Sandene, Bennett, Braswell, and Oranje (2005, p. xi).

## Endnotes

- 1 A fourth-grade sample was also tested on computer. See Sandene, Bennett, Braswell, and Oranje (2005) for a description of the sample, analyses and results.
- 2 Percentages of schools and students are weighted and may differ substantially from raw percentages.
- 3 One possible cause of these differences is the extent to which students omit, don't reach, or give off-task responses more frequently in one versus the other mode. Analysis of the percentages of students giving such responses suggested that the differences were so small as to be of limited consequence. (See Sandene, Bennett, Braswell, and Oranje [2005, Appendix C] for details of these results.)
- 4 IRT  $b$  parameter estimates can also be compared across paper forms administered to different nationally representative samples as part of the MOL study. For this comparison, the points clustered evenly around the identity line, further evidence that the difference in difficulty apparent in Figure 1 is indeed an effect of the mode of presentation and not simply variation due to the examinee sample. See Sandene, Bennett, Braswell, and Oranje (2005, pg. 15) for the resulting scatter plot.
- 5 Item difficulty was also evaluated in the proportion-correct ( $p+$ ) metric. Over all items, the median of the difficulty differences was  $-.05$  (range =  $-.17$  to  $.02$ ). The median difference for the items needing considerable change was  $-.08$  and the median difference for the items needing minimal change was  $-.04$ . With regard to item format, the median difference for the short- and extended-constructed-response items was  $-.08$ , whereas the comparable value for multiple-choice items was  $-.03$ . Thus, in general, the  $p+$  results are consistent with the differences in the  $b$  parameter estimates.
- 6 Several multiple-choice items also appeared to show large mode effects (e.g., #6, 21, 24). These effects might have been caused by a variety of factors. For example, students taking the paper test might have used their hand-held calculators more frequently or more effectively (for items 21 and 24) than MOL students used the online calculator, or students may have worked problems more frequently or more thoroughly in their paper booklets than students made use of scratch paper in the online condition.

## Author Biographies

Randy Elliot Bennett is Distinguished Scientist in the Research & Development Division at Educational Testing Service in Princeton, New Jersey. A graduate of Teachers College, Columbia University, Dr. Bennett began his employment at ETS in 1979. Since the 1980s, he has conducted research on the applications of technology to testing, on new forms of assessment, and on the assessment of students with disabilities. Dr. Bennett's work on the use of new technology to improve assessment has included research on presenting and scoring open-ended test items on the computer, on multimedia and simulation in testing, and on generating test items automatically. For this work, he was given the ETS Senior Scientist Award in 1996 and the ETS Career Achievement Award in 2005. He is the author of many publications including "Technology and Testing" (with Fritz Drasgow and Ric Luecht) in *Educational Measurement* (4th Edition) and "What Does it Mean to Be a Nonprofit Educational Measurement Organization in the 21st Century" (<http://www.ets.org/Media/Research/pdf/Nonprofit.pdf>). He can be contacted at [rbennett@ets.org](mailto:rbennett@ets.org).

James S. Braswell is a senior assessment development specialist at the American Institutes for Research in Washington, DC. From 1969 to 2004 he was employed by Educational Testing Service in Princeton, NJ where he developed a variety of mathematics examinations. He received his bachelor of arts degree in mathematics from the University of Tennessee, a master's degree in mathematics from George Peabody College of Vanderbilt University, and a PhD in curriculum and instruction from the University of Wisconsin, Madison. He has worked with others to develop mathematics examinations for programs such as the SAT, GRE, NAEP, and the computer-delivered component of the Math On-Line study. He has also worked on the development of mathematics assessments for several state testing programs. He can be contacted at [jbraswell@air.org](mailto:jbraswell@air.org).

Andreas Oranje is a Psychometric Manager at Educational Testing Service in Princeton, NJ. He received his Ph.D. in psychology from the Department of Psychological Methods at the University of Amsterdam, the Netherlands, in 2001 and his Masters Cum Laude from the same department in 1998. His interests cover various psychometric and statistical topics such as (hierarchical) latent variable modeling, variance estimation, and data visualization. His experiences include design and analysis of the National Assessment of Educational Progress, various applications of structural equation

models with categorical variables, and statistical inference with complex samples. He can be contacted at [aorange@ets.org](mailto:aorange@ets.org).

Brent Sandene is an Assessment Specialist at Educational Testing Service in Princeton, NJ, where he has been employed since 1994. His degrees include a B.Music degree from the University of Nebraska-Lincoln, M.Mus.Ed. from Indiana University, and a Ph.D. in Music Education from the University of Michigan. His interests include item development for tests in the arts and the uses of technology for standardized testing. He has worked on projects such as item development for the Praxis series of tests in Music, the Major Field Test in Music and the Advanced Placement Music Theory Examination, and served as coordinator of the NAEP Assessments in Music. He can be contacted at [bsandene@ets.org](mailto:bsandene@ets.org).

Bruce A. Kaplan holds the title of director of data analysis and interactive systems in the Research & Development Division at Educational Testing Service, in Princeton, NJ. He has been employed at ETS since 1979. He received his master's degree in economic and social statistics from the Industrial and Labor Relations School at Cornell University in 1979. He received his bachelor of science with highest of honors in applied mathematics and computer science from the State University of New York at Stony Brook in 1976. His interests include sample survey design and estimation, exploratory data analysis, regression analysis, computerization of statistical techniques, empirical Bayes techniques, and graphical displays. His experiences cover a wide range of research, computer technology, and statistical areas. He can be contacted at [bkaplan@ets.org](mailto:bkaplan@ets.org).

Fred Yan holds the title of research data analyst in the Center for Data Analysis Research in the Research and Development Division at Educational Testing Service, in Princeton, NJ. He has been employed at ETS since 1998. He received his bachelor's degree in Engineering Mechanics from Tianjin University, China and his master of science degree in Civil Engineering from Pennsylvania State University. He also is currently working toward completing the requirements for the master of science degree in computer science from New Jersey Institute of Technology. His experience covers statistical data analysis and application software development. He can be contacted at [fyan@ets.org](mailto:fyan@ets.org).

## Appendix A

### Notes on Sampling Methodology

The primary sampling units (PSUs) were counties or groups of counties. Because the study did not require the same large sample sizes as a NAEP assessment, a subset of 52 PSUs was sampled from the 94 PSUs selected for the NAEP 2001 history and geography assessments (Lapp, Grigg, & Tay-Lim, 2002; Weiss, Lutkus, Hildebrant, & Johnson, 2002). To increase the chance of getting a representative subset, the sampling was constrained to include the 10 largest PSUs, half of the 12 smallest PSUs, and half of the remaining 72 PSUs.

Middle and secondary schools were chosen (without replacement) across all selected PSUs from a sorted list, with probabilities proportional to size. The sample was designed to over-sample large schools and schools with more than 10 percent Black students or 10 percent Hispanic students.

Students who were judged by standard NAEP exclusion criteria as not being able to participate meaningfully in the testing activities without accommodations were excluded. Ninety-four of the 1,297 sampled students were excluded from online testing and 229 of 3,522 sampled students were excused from paper testing. These exclusion rates are similar to those for non-accommodated samples tested in the NAEP 2001 assessments in history and geography (Lapp, Grigg, & Tay-Lim, 2002; Weiss, Lutkus, Hildebrant, & Johnson, 2002).

## Appendix B

**Table 1B: Inter-rater Reliability for Constructed-Response Items**

Item	Percentage Exact Agreement	
	P&P	MOL
2	99	98
3	95	92
7	93	91
10	80	84
13	99	98
15	97	98
16	99	98
17	99	98
19	94	90
26	85	85

NOTE: MOL = Math Online. The number of students responding ranged from 239 to 254 on the paper test; from 249 to 253 on MOL Items. Items 10 and 26 were scored on a 5-point scale. All other items were scored on 2- or 3-point scales.

## Appendix C

### Questions and Number of Response Categories for the Computer Experience Measure

- How often do you use a computer at school? (5)
- How often do you use a computer outside of school? (5)
- Is there a computer at home that you use? (2)
- Do you use the Internet at home? (2)
- To what extent do you do the following on a computer?
  - Play computer games (4)
  - Write using a word processing program (4)
  - Make drawings or art projects on the computer (4)
  - Make tables, charts, or graphs on the computer (4)
  - Look up information on a CD (4)
  - Find information on the Internet for a school project or report (4)
  - Find information on the Internet for personal use (4)
  - Use e-mail to communicate with others (4)
  - Talk in chat groups or with other people who are logged on at the same time you are (4)
- When you do mathematics in school, how often do you do each of the following?
  - Use a computer (4)

## The Journal of Technology, Learning, and Assessment

### Editorial Board

**Michael Russell, Editor**  
Boston College

**Allan Collins**  
Northwestern University

**Cathleen Norris**  
University of North Texas

**Edys S. Quellmalz**  
SRI International

**Elliot Soloway**  
University of Michigan

**George Madaus**  
Boston College

**Gerald A. Tindal**  
University of Oregon

**James Pellegrino**  
University of Illinois at Chicago

**Katherine Bielaczyc**  
Museum of Science, Boston

**Larry Cuban**  
Stanford University

**Lawrence M. Rudner**  
Graduate Management  
Admission Council

**Marshall S. Smith**  
Stanford University

**Paul Holland**  
Educational Testing Service

**Randy Elliot Bennett**  
Educational Testing Service

**Robert Dolan**  
Pearson Education

**Robert J. Mislevy**  
University of Maryland

**Ronald H. Stevens**  
UCLA

**Seymour A. Papert**  
MIT

**Terry P. Vendlinski**  
UCLA

**Walt Haney**  
Boston College

**Walter F. Heinecke**  
University of Virginia